Huidige vuistregels zijn opgesteld voor vraagselectie bij de constructie van een gestandaardiseerd instrument in een normgeoriënteerde toetssituatie.

Valide vuistregels voor de evaluatie van studietoetsen: Een oproep tot verbetering

Niels Smits
Dr. N. Smits is als universitair hoofddocent werkzaam bij het Research Institute of Child Development and Education, Faculteit der Maatschappij- en Gedragswetenschappen, Universiteit van Amsterdam en is als toetsexpert verbonden aan de lokale examencommissie. E-mail: n.smits@uva.nl.

Sharon Klinkenberg
Dr. S. Klinkenberg is als docent werkzaam bij Communicatiewetenschap en is co-directeur van het Teaching & Learning Center van de Faculteit der Maatschappij- en Gedragswetenschappen, Universiteit van Amsterdam; E-mail: s.klinkenberg@uva.nl.

In ons eerdere artikel ‘Gangbare vuistregels voor indicatoren van toetskwaliteit zijn te strikt’ (Smits & Klinkenberg, 2026) lieten wij, onder andere door middel van een simulatiestudie, zien dat veelgebruikte normen voor toetskwaliteit, zoals die in het handboek van Van Berkel en Bax (2023, zie Tabel 1) en in toetssoftware, vaak onnodig streng zijn. Daardoor kunnen vragen of hele tentamens ten onrechte als onvoldoende worden beoordeeld, terwijl ze in de praktijk goed bruikbaar zijn. In dit artikel bouwen wij voort op die bevindingen en vertalen we ze naar de dagelijkse toetspraktijk. Aan de hand van een concrete tentamenanalyse laten we zien wat er gebeurt als je een tentamen beoordeelt met de gangbare vuistregels en hoe het beeld verandert als je rekening houdt met enkele basale principes die bekend zijn uit de testleer. We besluiten met praktische aanbevelingen voor realistischere normen voor tentamenanalyses.

Tabel 1. Een selectie van vuistregels voor statistische indicatoren volgens Van Berkel en Bax (2023).

p-waarde				Onderscheidingsvermogen		Cronbach’s alfa	
Soort vraag	onder	Ideaal	boven	goed/zeer goed	0.35 en hoger	goed/zeer goed	0.90 en hoger
open	0.25	0.50	0.90	voldoende/goed	0.25-0.35	voldoende/goed	0.80-0.90
vierkeuze	0.44	0.62	0.90	middelmatig/ voldoende	0.15-0.25	middelmatig/ voldoende	0.70-0.80
				slecht/middelmatig	minder dan 0.15	slecht/middelmatig	minder dan 0.70

De kritiek in het kort

De kritiek was dat de huidige vuistregels oorspronkelijk zijn opgesteld voor vraagselectie bij de constructie van een gestandaardiseerd instrument in een normgeoriënteerde toetssituatie. Zowel taak, toetsdoel als soort meetinstrument sluit niet aan bij de praktijk van het Hoger Onderwijs. Daar selecteert men niet de beste vragen voor toekomstige afnames, maar wordt gecontroleerd of reeds gebruikte toetsvragen goed genoeg waren; wordt de toets niet gebruikt om prestaties van studenten met elkaar, maar met een onderwijsdoelstelling te vergelijken; en vormt het beoogde instrument geen geijkte schaal maar een studietoets. Deze mismatch zorgt voor een groot risico op het afkeuren van geschikte vragen (en daardoor op slechtere toetsen). Onze simulaties bevestigden dit: hoewel alle toetsen even moeilijk en onderscheidend waren, gaven de indicatoren systematische vertekeningen bij vragen met veel variatie in moeilijkheid en studentgroepen met weinig individuele verschillen. Daarnaast toonden we aan dat criteriumgeoriënteerd toetsen anderssoortige kwaliteitsindicatoren vereist.

De toetsevaluatie

Het tentamen Onderzoeksmethodologie werd in oktober 2025 afgenomen bij 45 studenten van de Universitaire Pabo van Amsterdam (UPvA). Het bestond uit 29 vierkeuzevragen en één open vraag over een breed scala aan onderwerpen als onderzoeksdesigns en steekproeftrekken en had betrekking op kennisniveau's onthouden, begrijpen en toepassen. Elke vraag leverde 0 of 1 punt op en de cesuurscore voor een voldoende was 19 goed. De geschatte standaardafwijking en Cronbach's alfa waren, respectievelijk, 3.43 en 0.56, de resulterende standaardmeetfout 2.26. In Tabel 2 staan voor elke vraag schattingen van de indicatoren van toetskwaliteit.

Er was veel spreiding in beide indicatoren: De moeilijkheid, uitgedrukt in de 'p-waarde', liep van 0.27 tot 0.98. Merk op dat de standaarddeviatie (SD) direct volgt uit de p-waarde: Hoe dichter deze bij 0.5 ligt, hoe groter de SD (en dus hoe meer studenten onderling verschillen). Het onderscheidingsvermogen, uitgedrukt in de correlatie tussen de itemscore en het aantal goede antwoorden op de hele toets (Rit), varieerde van -0.05 tot 0.55.
Als we de regels uit Tabel 1 toepassen is de p-waarde van negen opgaven niet ideaal, vijf zijn te hoog en vier te laag. Weliswaar zijn de Rit-waarden van, respectievelijk, tien, negen en drie opgaven zeer goed, goed en voldoende, maar de overige acht opgaven zijn slecht. Tenslotte is de geschatte betrouwbaarheid (alfa was 0.56) van de toets slecht. Het is begrijpelijk dat docenten zich afvragen of de toets moet worden aangepast.

We introduceerden in onze publicatie vier vragen die men kan stellen om te bepalen hoe strikt de huidige vuistregels genomen moeten worden:

1. *Wat is de verklaring voor een afwijkende waarde?* Lage indicatorwaarden mogen op zichzelf geen reden zijn om de toets aan te passen, er moet altijd extra informatie worden verzameld om te kijken wat een mogelijke verklaring is. Van alle vragen met een afwijkende p- en/of Rit-waarde werden de inhoud, bijbehorende studiestof en collegeslides geïnspecteerd, maar nergens kon een tekortkoming worden gevonden.

2. *Wat is het toetsdoel?* Bij normgeoriënteerd toetsen staat het onderling vergelijken van studentprestaties centraal, bij criteriumgeoriënteerd toetsen of studenten voldoen aan minimale vereisten. Wij achtten het tweede doel het belangrijkst, maar de vuistregels uit Tabel 1 zijn opgesteld voor het eerste. Daar waar voor normgeorienteerde situaties p-waarden dicht bij 0.5 gewenst zijn, geeft de testleer voor criteriumgeoriënteerde settings geen algemene regels; ze volgen direct uit het kennisdomein en in dit geval sloot volgens de domeinexpert de moeilijkheid redelijk aan op die van de betreffende stof en vereiste kennisniveau's. Indien er voldoende

Tabel 2. Itemstatistieken van Tentamen Onderzoeksmethodologie.

	Moeilijkheid		Onderscheidingsvermogen		
Vraag	P	SD	Rit	onder	boven
1	0.84	0.37	0.37	0.06	0.68
2	0.62	0.49	0.48	0.25	0.71
3	0.69	0.47	0.42	0.19	0.65
4	0.87	0.34	**0.04**	-0.12	0.20
5	0.73	0.45	**0.06**	-0.27	0.39
6	**0.98**	0.15	0.24	0.01	0.48
7	**0.91**	0.29	0.32	-0.01	0.65
8	0.78	0.42	0.27	-0.04	0.57
9	0.89	0.32	**0.11**	-0.20	0.43
10	0.33	0.48	0.25	0.01	0.49
11	**0.38**	0.49	**-0.05**	-0.34	0.25
12	**0.53**	0.50	0.39	0.17	0.60
13	0.71	0.46	**0.08**	-0.18	0.34
14	**0.27**	0.45	0.35	0.12	0.58
15	**0.93**	0.25	0.25	-0.08	0.58
16	**0.93**	0.25	0.30	-0.01	0.61
17	0.64	0.48	0.42	0.20	0.65
18	0.47	0.50	0.48	0.30	0.66
19	0.47	0.50	0.33	0.09	0.57
20	0.82	0.39	**0.07**	-0.16	0.29
21	0.53	0.50	0.15	-0.13	0.43
22	0.73	0.45	0.35	0.11	0.60
23	0.71	0.46	0.55	0.32	0.78
24	0.62	0.49	**-0.02**	-0.28	0.25
25	0.73	0.45	0.49	0.24	0.73
26	0.71	0.46	**0.11**	-0.24	0.46
27	0.80	0.40	0.35	0.05	0.65
28	**0.38**	0.49	0.43	0.25	0.62
29	0.53	0.50	0.30	0.01	0.58
30	**0.96**	0.21	0.35	-0.07	0.77

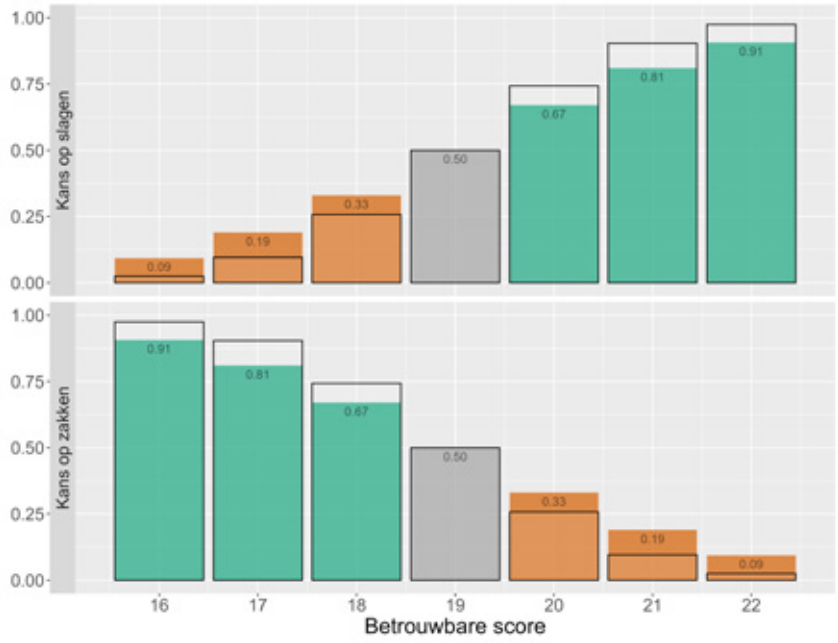
Noot: P is proportie correct ('p-waarde'), SD is standaardafwijking, Rit is item-testcorrelatie en onder en boven zijn, respectievelijk, de onder- en bovengrens van het 95%-betrouwbaarheidsinterval van Rit; **Dikgedrukte** getallen wijken volgens de vuistregels van Tabel 1 af.

spreiding in de scores is, kunnen bij criteriumgeoriënteerd toetsen de populaire schattingen van de betrouwbaarheid, zoals alfa, toch worden gebruikt, maar dan zijn de standaarden uit Tabel 1 te streng: een geschatte betrouwbaarheid zoals die van de huidige toets is dan namelijk niet slecht (Ebel & Frisbie, 1991). Maar criteriumgeoriënteerd toetsen vraagt ook om extra analyses (zie hieronder).

3. *Wat is het type instrument?* Standaardisatie betreft het vaststellen van een geijkte schaal met normen voor de testpopulatie. De ontwikkeling van gestandaardiseerde instrumenten bevat meerdere rondes van

pilotafnames en aanpassingen en vindt vooral plaats bij instellingen met voldoende middelen zoals het Cito, testuitgevers of wetenschappelijke instituten. Alleen de allerbeste vragen resteren en het is niet erg als een toereikende vraag afvalt. De toetspraktijk van het Hoger Onderwijs is heel anders (Standards, 2014, p. 183): Docenten zijn meestal door het examenregelement gedwongen om toetsen met nieuwgemaakte vragen te gebruiken en voeren geen selectie maar een kwaliteitscontrole uit. Het ten onrechte verwijderen van een geschikte vraag heeft veel impact omdat het de toetskwaliteit juist kan verslechteren. Door hun gevoeligheid voor onderschatting van toetskwaliteit zijn de vuistregels uit Tabel 1 dus ongeschikt voor kwaliteitscontroles. Bovendien bieden de ondergrenzen voor voldoende kwaliteit mogelijk enig houvast, maar kwalificaties als '(zeer) goed' zijn irrelevant.

4. *Hoe zit het met de spreiding?* We vonden dus lage waarden op de indicatoren, maar daar zijn drie verklaringen voor die te maken hebben met de spreiding van de scores. De eerste verklaring is dat bij lage spreiding in vaardigheid de indicatoren een systematische onderschatting geven van toets- en vraagkwaliteit. We weten dat UPvA-studenten, in vergelijking met andere studentgroepen, doorgaans niet veel spreiding laten zien. De tweede verklaring is dat in situaties met veel spreiding in vraagmoeilijkheid, de indicatoren de toetskwaliteit systematisch onderschatten. Bij de constuctie van de huidige toets werd een criteriumgeoriënteerd doel nagestreefd wat resulteerde in een wisselende



Figuur 1. Voor een reeks betrouwbare scores de kans op slagen (boven) en zakken (onder) bij de huidige SEM. Groene en rode kleurvakken geven, respectievelijk, kansen op goede en foute beslissingen; de zwarte rechthoeken zijn referentiewaarden voor een SEM onder een 'goede' betrouwbaarheid (0.80).

vraagmoeilijkheid. Een derde verklaring vormt het kleine aantal studenten (45). Met name correlaties zijn gevoelig voor steekproeffluctuaties en daarom zijn in Tabel 2 ook de onder- en bovengrens van betrouwbaarheidsintervallen geplaatst, die aannemelijke waarden van de werkelijke Rit's geven. Het bleek dat voor alle gemarkeerde vragen het interval de ondergrens voor voldoende waarden (0.15) bevatte. Dezelfde aanpak voor alfa gaf het betrouwbaarheidsinterval (0.34, 0.79) waarin de ondergrens voor voldoende betrouwbaarheid (0.70) een aannemelijke waarde was. Dus als er rekening wordt gehouden met de spreiding lijkt de kwaliteit van de toets zo slecht nog niet.

Ook werd gesteld dat bij criteriumgeoriënteerd toetsen het belangrijk is wat de gevolgen van meetfouten zijn voor zak/slaagbeslissingen. Op basis van de geschatte standaardmeetfout (SEM) kan worden berekend hoe groot de kans is dat een student terecht of onterecht slaagt of zakt. In Figuur 1 zijn deze kansen voor verschillende 'ware' of betrouwbare scores (Drenth & Sijtsma, 2006) geschat. Goede beslissingen zijn hierin groen weergegeven en foute beslissingen rood. Neem bijvoorbeeld een betrouwbare score van 17. Deze ligt onder de cesuur, dus de student zou eigenlijk moeten zakken. Door meetfout kan de geobserveerde score soms toch boven de cesuur uitkomen. In dit geval is de kans op een goede beslissing (de student zakt) 0.81. De kans op een foute beslissing (een student die eigenlijk zou moeten zakken maar toch slaagt) is 0.19. Aan de andere kant van de cesuur speelt het omgekeerde. Bij een betrouwbare score van 20 zou de student eigenlijk moeten slagen. Toch kan de geobserveerde score door meetfout onder de cesuur vallen. In dit geval is de kans op een goede beslissing 0.67 en de kans op een foute beslissing (een student die ten onrechte zakt) 0.33. Uit de figuur blijkt dat hoe verder de betrouwbare score van de cesuur af ligt, hoe kleiner de kans op een verkeerde beslissing wordt. De grootste kans op fouten

Docenten moeten zich afvragen of de toets op basis van de analyse moet worden aangepast.

ligt altijd vlak rond de cesuur. In dit voorbeeld is de maximale kans op een fout 0.33, wat als acceptabel kan worden beschouwd. Ten slotte is het goed om te beseffen dat Cronbach's alfa een ondergrens van de betrouwbaarheid is (Drenth & Sijtsma, 2006). De werkelijke meetfout zal dus waarschijnlijk iets kleiner zijn, waardoor de werkelijke kansen op foute beslissingen naar verwachting nog iets lager liggen.

Conclusie

Populaire vuistregels voor indicatoren van toetskwaliteit zijn vaak ongeschikt voor tentamens in het Hoger Onderwijs. Zoals geïllustreerd met de analyse van een tentamen Onderzoeksmethodologie, kun je door vier aanvullende vragen te stellen en een extra analyse uit te voeren beter inzicht krijgen in de kwaliteit van een tentamen. Het tentamen Onderzoeksmethodologie bleek veel beter te functioneren dan de vuistregels

deden vermoeden.
Het grootste probleem is dat de vuistregels uitgaan van populaties met grote individuele verschillen en items met soortgelijke inhoud, terwijl studietoetsen juist vaak worden afgenomen in groepen met kleine onderlinge verschillen (bijvoorbeeld studenten in het eerste bachelorjaar) en gemengde inhoud (bijvoorbeeld opgaven over verschillende boekhoofdstukken), waardoor de toetskwaliteit systematisch onderschat wordt. Ook zijn de groepen vaak klein waardoor de impact van steekproeffluctuaties groot is. Kwaliteitscontrole is een belangrijke stap in het toetsproces en het is essentieel dat daarbij valide normen worden gehanteerd. We roepen hierbij dan ook op tot een afzwakking van grenswaarden en genuanceerder gebruik van (aanvullende) indicatoren. Meer specifiek willen we auteurs van handboeken, zoals Van Berkel en Bax, en ontwikkelaars van toetssoftware uitnodigen om onze bevindingen en aanbevelingen te overwegen bij toekomstige herzieningen. ■

Referenties

- Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen* (4de dr.). Bohn Stafleu Van Loghum.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5de dr.). Prentice-Hall.
- Smits, N., & Klinkenberg, S. (2026). Gangbare vuistregels voor indicatoren van toetskwaliteit zijn te strikt. Tijdschrift Voor Hoger Onderwijs, 44.
- Standards. (2014). *Standards for educational and psychological tests* [American Psychological Association and American Educational Research Association and National Council on Measurement in Education]. American Psychological Association.
- Van Berkel, H., & Bax, A. (2023). Het meten van psychometrische toetskwaliteit. In H. Van Berkel, A. Bax, D. Joosten-ten Brinke, K. Beekman, & T. van Schilt-Mol (Red.), *Toetsen in het hoger onderwijs* (5de dr., pp. 81-94). Bohn Stafleu Van Loghum.

Gesignaleerd

U.S. Education News
'A' Grades Are Suddenly Everywhere Since the Arrival of ChatGPT
AI is accelerating grade inflation, research indicates, and making it harder for employers to size up graduates

By Lindsay Ellis, May 13, 2026 7:00 pm ET
AI is making "A" grades easier to come by, a new study shows—and making them less useful to employers trying to size up college graduates.
The share of A's in college classes heavy on writing and coding—in other words, work more prone to artificial intelligence use—has grown more significantly than in other classes since ChatGPT's debut, according to a paper from the University of California, Berkeley, released Wednesday. Professors teaching AI-exposed classes gave out about 30% more A's and fewer A-minus and B-plus grades.
Lees verder op: https://www.wsj.com/us-news/education/a-grades-are-suddenly-everywhere-since-the-arrival-of-chatgpt-845baae7?mod=WTRN_pos1

onderdeel van CITO

Blogserie

Het is weer examentijd in het vo! Benieuwd hoe wij daaraan bijdragen?

Lees hier onze blogs →

We moeten rekening houden met al die vakken, richtingen, niveaus en docenten die niet één homogene groep vormen.

Chiel Huijskes, *beleidsadviseur Strategie Beleid en Vernieuwing*

'We vragen docenten geregeld naar hun ideeën over digitalisering en naar wat zij en hun leerlingen belangrijk vinden. Die input uit de praktijk is heel waardevol.'

Stephanie Kruiper, *onderwijskundige programmateam DCE*

CITO CTE